Collect Ethically: Reduce Bias in Twitter Datasets.

Lulwah Alkulaib, Abdulaziz Alhamadani, Taoran Ji, and Chang-Tien Lu

Virginia Tech, Falls Church, VA 22043, USA
{lalkulaib, hamdani, jtr, ctlu}@vt.edu

Abstract. The Twitter platform is appealing to researchers due to the ease of obtaining data and the ability to analyze and produce results rapidly. However, sampling Twitter data for research purposes needs to be regulated to produce unbiased results. In this paper, factors that lead to sampling bias are addressed, case studies that have been encountered are presented, and an approach is proposed to reduce sampling bias and flaws in datasets collected from Twitter. Then, experiments are conducted on two case studies, and a larger dataset is achieved by following the proposed guideline. The results indicate that using multiple Twitter application programming interfaces (APIs) for data collection is the best way to obtain a randomly sampled dataset.

Keywords: Sampling Bias · Twitter dataset · Twitter API.

1 Introduction

Twitter allows its users to publish 'tweets'up to 280 characters long that are visible to other users in their network. Users that tweet publicly generate a huge amount of data that is available for collection by researchers if needed. Ahmad et. al.[9] explain how Twitter is one of the most researched platforms due to data availability, the ease of following conversations, and the ability to use hashtags as a topic-grouping mechanism. However, collecting Twitter data without any considerations will result in a biased dataset. Using the Twitter developer platform, researchers can collect and analyze tweets using one of three application programming interfaces (APIs): Twitter Search API, Twitter Streaming API, or Twitter Firehose. As noted in Table 1, Twitter data access varies depending on the chosen access method. This casts doubt on conclusions drawn from Twitter data samples and impacts the reliability of research findings based on this data.

In research, any trend or deviation from the truth in data collection, analysis, interpretation, or publication is called bias. In our research, the study focuses on sampling bias, which occurs when a sample is collected in such a way that some members of the intended population are not equally represented in the sample, resulting in a non-random sample. If such an error occurs in sampling, the results of the research could be mistakenly attributed to the study phenomenon instead of the sampling method[2]. This leads us to question the use of Twitter as a data source while considering whether sampling bias has occurred in surveys

and its effect on the resulting conclusions. To address this issue, a Twitter data collection guideline is proposed. The main contributions of this paper are as follows:

- Formulating a query expansion guideline consisting of six factors, which when followed, minimizes sampling bias when using Twitter as a data source.
- Using three Twitter APIs (i.e., search, streaming, and firehose) as methods of data collection. The new datasets are collected by following the original query and the expanded query using the proposed guideline. Then, the results of each case study are compared.
- Conducting extensive experiments to validate the proposed guideline's effectiveness and comparing the expanded query results with non-expanded results.

2 Related work

The availability of Twitter data via APIs has made researchers eager to utilize them in studies. Below, a review of previous studies involving Twitter APIs presented.

Streaming API vs. Firehose: Wang et. al.[8] performed a comparative analysis of data samples obtained using Twitter Streaming API and studying the sampling bias that occurs in each sample. They concluded that using the Streaming API does not present a sample as representative as the Firehose API. **Search API vs. Streaming API:** Filho et al.[3] discussed population sample bias in Twitter data and how it impacts predictive accuracy. They studied the available Twitter data and whether it is comprehensive enough to make user characterizations or predict outcomes. They found that using free Twitter APIs to collect samples is not sufficient to make accurate predictions. **Search API vs. Firehose:** Taking a different approach, Zhang et. al.[10] compared three collection methods in event studies: keyword filtering using the Search API, geolocation filtering using the Search API, and random sampling using the Twitter Firehose API. They attributed keyword sampling bias due to outcome selection and proposed geolocated and random sampling as a solution to reduce bias.**Search API vs. Streaming API vs. Random Firehose Sample:** In 2017, Morstatter et al.[4] addressed detecting bias from sampling strategies while comparing datasets with unsampled Twitter Firehose data. They presented multiple strategies to mitigate bias using the Twitter Streaming API. **Hashtag Sampling Bias:** Morstatter et al.[5] measured sampling bias differently. Their work focused on hashtags and their representativeness in the sample in comparison to trends on Twitter. A hashtag is 'biased' if the relative trend is overrepresented or underrepresented to a statistically significant degree when compared to its true trend on Twitter.

To the best of our knowledge, none of the previous studies presented a comparison of the main three Twitter APIs that researchers use as a data collection tool. A new guideline is proposed and applied by collecting and comparing datasets using Twitter Search, Streaming, and Firehose APIs.

Table 1: Twitter API Comparison

API	No Query	Query				Results	Free
		Keyword	Username	Location	By percentage	By number	
Search API	×	✓	✓	×	×	✓	✓
Streaming API	×	✓	✓	✓	✓	×	✓
Firehose	×	✓	✓	✓	✓	×	×

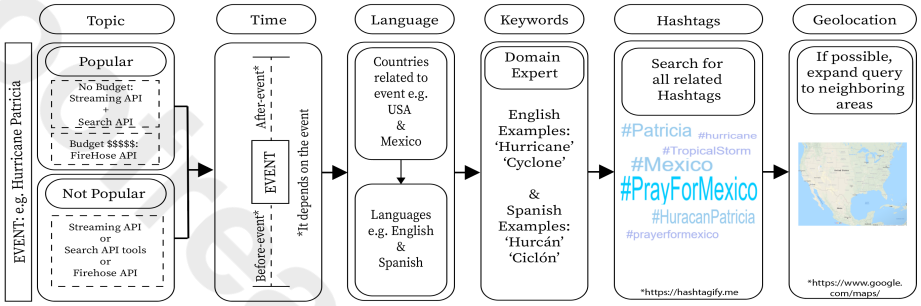


Fig. 1: The flowchart of the proposed data collection approach

3 Proposed Approaches

Data Collection Guideline: The purpose of this work is to build a guideline that will serve as a reference for researchers who collect data from Twitter. The guideline will help minimize sampling bias in Twitter-based research datasets; see (Fig. 1). There are six aspects that researchers should consider since they affect the Twitter APIs' performance.

Topic: Knowing a topic's popularity is essential since the Twitter Streaming API does not capture more than 40% of real-time tweets during popular events[1]. For example, during presidential elections, using the Streaming API alone would lead to a biased dataset. However, if the topic is unpopular (i.e., fewer than six tweets/minute are related to the topic[6]), Twitter Streaming API would be sufficient for data collection.

Time: The key to selecting the data collection period is considering the longest related date range. Collecting data during an event is the most common method. However, events differ in how users tweet about them. Event-related tweets can occur before, during, and after an event or just during and after an event. It is tempting to only collect data during an event, but as shown in our study, a significant number of tweets will be missed when the time factor is not considered. Therefore, collecting Twitter data before and after an event would result in a more inclusive dataset.

Language: When writing an API query, it is crucial to select the language of returned tweets. If the selected event includes different languages, then all lan-

guages should be included in the query. For example, in our first case study, Hurricane Patricia traveled through Mexico and the United States (US). Therefore, tweets written in Spanish and English were included in the research.

Keywords: Keyword expansion should be performed by a domain expert. It is important to know what other related words could be used for a certain research topic, and using the expertise of a person with specialized knowledge of the domain will help expand the query correctly.

Hashtags: Expanding the query by utilizing related hashtag topics is essential. For example, when researching Hurricane Patricia, we searched for related hashtags at[1] *(see Fig. 1 Hashtags' column).*

Geolocation: Finally, if the location of the initial topic spread is known, expanding the query by location into neighboring areas should also be considered. Using this guideline as a reference when collecting data from Twitter should ensure that there is a standardized method for data collection. It also allows researchers to verify that they avoided factors that could cause sampling bias.

3.1 Methods

In this section, each API's data collection method is described. *Streaming API:* Two datasets collected via the Streaming API and published[11] were chosen. Since published Twitter datasets are only allowed to publish tweet identifications (IDs) to be compliant with Twitter Terms of Service (TOS), the datasets are hydrated[7] using Hydrator[2]. In the dataset description, it is mentioned that the data are collected using the Streaming API and filtered to only capture English results. *Search API:* Twitter's Search API only allows access to data covering the last seven days. Since both datasets covered from 2015 and 2016, a tool called 'GetOldTweets' [3] that bypasses the time limitation of the Search API was used. Then, the datasets were collected using the tool once with the exact query used by Zubiaga and then using the new expanded query after following the data collection guideline. *Firehose:* The Discovery Analytics Center (DAC) at Virginia Tech purchased 10% of the worldwide data on the Twitter Firehose between August 2014 and April 2018. Gnip, an API that allows access to Firehose data and filters it to form a subset of results, was used to access and collect the dataset, following the exact query used by Zubiaga, and then using our expanded query.

4 Experiments

Two case studies are assumed to have sampling bias due to their collection method. Below, we show how our proposed method is used to reproduce datasets and minimize bias.

[1] www.hashtagify.com
[2] https://github.com/DocNow/hydrator
[3] https://github.com/Jefferson-Henrique/GetOldTweets-python

Table 2: Chosen datasets.

Datatset	Hurricane Patricia	Egypt Air Hijacking
Date	October 24-December 8, 2015	March 29-30, 2016
Query	hurricane, patricia, #hurricanepatricia, #huracanpatricia	#egyptair, hijacked, plane, cyprus, airport
# of tweets	1,151,220	702,586

Table 3: Query expansion for event "Hurricane Patricia"

Date	2015-10-13 ∼ 2016-01-02
Keywords	#HuracanPatricia, #hurricanepatricia, #huracanpatricia, #globalwarming, #climatechange, #houstonflood, #hurricane, #prayformexico, hurricane, typhoon patricia, cyclone, Tropical storm, hurricane patricia, huracn, ciclon, tormenta tropical, tifon patricia

Table 4: Query expansion for event "Egypt Air Hijacking"

Date	2016-03-28 ∼ 2016-04-30
Keywords	#egyptair, #egyptairhijack, #الطائرة المصرية, #اختطاف طائرة مصرية, #hijacked, #Αεροπειρατεια, #MS181, #cyprushijack, Egypt Air, Hijack, MS181, Seif AlDin Mustafa, Hijacked, Hijacker, خاطف الطائرة المصرية, اختطاف الطائرة المصرية, الخطوط المصرية, الطياره المصريه, الطائرة المصرية, الطائره المخطوفه, سيف الدين مصطفى, الطيران المصري, الخطوط الجويه المصرية, المختطفة, الظائرة المصرية

4.1 Experiments Setup

Two datasets published by Zubiaga were chosen. The datasets were collected for two events. The datasets are explained in detail in Table 2. The suggested guideline was followed to expand the queries. Both chosen datasets were popular topics, and previous work has shown that using the Streaming API alone will create sampling bias. It was determined that the time buffer for the Hurricane Patricia dataset should begin with the weather predictions that preceded the hurricane and end with news outlets' last reports about the hurricane's after-effects. However, for the Egypt Air Hijacking dataset, one day prior to the incident was added as a buffer, and it ended with the latest news found about the court proceedings. For Hurricane Patricia, the language was expanded to include Spanish due to the hurricane passing through Mexico and the US, and the language was expanded to include Arabic and Greek for the Egypt Air Hijacking. The hashtags were expanded using co-occurrence and tools that analyzed related hashtags. The keywords were expanded using domain experts in Spanish, Arabic, and Greek and added relevant English keywords using co-occurrence. In the initial datasets, the tweets were not from a significant number of locations, and for this reason, the queries were not expanded by location.

4.2 Data Collection

Streaming API: *Before Hydration:* The initial datasets collected by Zubiaga are described in Table 2. The datasets were downloaded as JSON files containing available matching tweet IDs using Hydrator [11]. *After Hydration:* After hydrating the datasets, 654,965 tweets were retrieved for the Hurricane Patricia event, which was 43% fewer than the number of total tweets. For the second event, 473,210 tweets were retrieved, which was a 33% decrease. The loss is attributed to deleted accounts or deleted tweets.

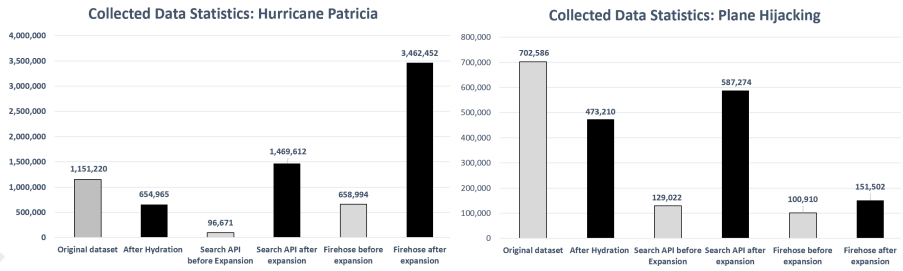
Search API: *Before Expansion:* GetOldTweets was used to collect data following the same dates and keywords presented in Table 2. The Hurricane Patricia event resulted in 96,671 tweets, whereas the Egypt Air Hijacking event resulted in 129,002 tweets. *After Expansion:* The queries of both events were expanded as shown in tables (3 and 4), respectively. The collected data for Hurricane Patricia using GetOldTweets resulted in 1,469,612 tweets, whereas 587,274 tweets were collected for the Egypt Air Hijacking event.

Firehose: *Before Expansion:* To be consistent, the available Firehose was used to collect data for both events following the same criteria presented in Table 2. The Hurricane Patricia event resulted in 658,994 tweets, and there were 100,910 tweets related to the Egypt Air Hijacking event. It is important to note that the low-return of Egypt Air Hijacking data is because the Firehose data in our lab was only 10% of the worldwide Twitter data. *After Expansion:* Following the expanded criteria in tables (3 and reftab3-2), the data for both events were collected from the available Firehose. The Hurricane Patricia and Egypt Air Hijacking events resulted in 3,462,452 and 151,502 tweets, respectively.

4.3 Results

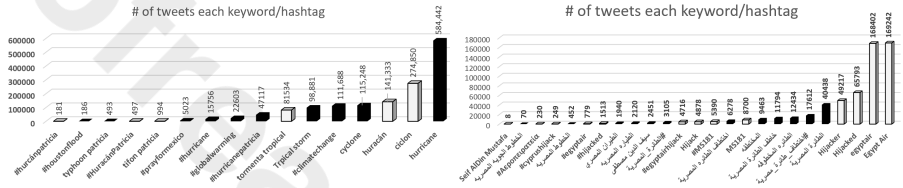
In this section, the impact of collecting data from multiple Twitter APIs is investigated for each case study.

Hurricane Patricia: As shown in (Fig. 2a), by comparing each API, it can be observed that the Streaming API data decreased 43%,the Search API data increased by more than 1400%, and the Firehose data increased by 425%. *Egypt Air Hijacking:* As depicted in (Fig. 2b), there was a 33% decrease in Streaming API-retrieved tweets, while Search API data increased by 355%, and the Firehose dataset increased by 50%. These results indicate that within the expanded timeframe, a larger number of relevant tweets were captured, which included some of the keywords, hashtags, or both. (Fig. 3a) and (Fig. 3b) show each keyword and hashtag used in the expanded query, along with their counts in the retrieved tweets. It is apparent that including words from different languages retrieved a lot of missing tweets, which allows the datasets to be representative of the population. In (Fig. 4a) and (Fig. 4b), the intersection between different APIs is investigated to study the number of overlapping tweets. It is evident that in events as popular as Hurricane Patricia, the overlap of data between APIs was



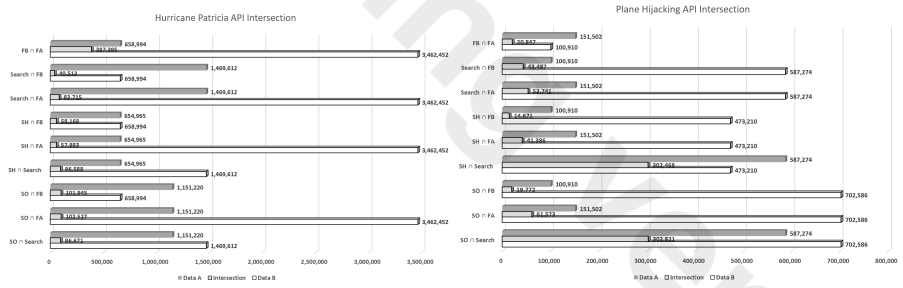
(a) Hurricane Particia. (b) Egypt Air Hijacking.

Fig. 2: APIs dataset comparison.



(a) Hurricane Particia. (b) Egypt Air Hijacking.

Fig. 3: Search API expanded query Keywords and Hashtags.



(a) Hurricane Particia. (b) Egypt Air Hijacking.

Fig. 4: APIs dataset intersection. *Legend: SO: Streaming API Originial; SH: Streaming API after Hydration; Search: Search API after expansion; FB: Firehose before Expansion; FA: Firehose after Expansion*

minimal. This leads us to conclude that using multiple Twitter APIs retrieves a larger number of relevant tweets, minimizing sampling bias.

4.4 Experiment Discussion

– The experiment finds that the representativeness of data collected from Twitter depends on both the query used and the API.

– Twitter Firehose is the only source that provides 100% of the tweets, but it comes at a cost that not all researchers can afford.
– To generate datasets with minimal bias using free Twitter APIs, combining multiple APIs results in a representative dataset. Additionally, expanding the search query could retrieve different tweets when querying multiple APIs.

5 Conclusion

This paper provides an analysis of Twitter data sampling obtained from the Twitter Streaming, Search, and Firehose APIs for two case studies. Comparing datasets collected using the original query and our guideline shows that using a single sampling method could lead to sampling bias. The experiments validate the effectiveness of the guideline and demonstrate growth in the resulting datasets when followed.

References

1. Campan, A., Atnafu, T., Truta, T.M., Nolan, J.: Is data collection through twitter streaming api useful for academic research? 2018 IEEE International Conference on Big Data (Big Data) (2018). https://doi.org/10.1109/bigdata.2018.8621898
2. Cortes, C., Mohri, M., Riley, M., Rostamizadeh, A.: Sample selection bias correction theory. In: International conference on algorithmic learning theory. pp. 38–53. Springer (2008)
3. Filho, R.M., Almeida, J.M., Pappa, G.L.: Twitter population sample bias and its impact on predictive outcomes. Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM 15 (2015). https://doi.org/10.1145/2808797.2809328
4. Morstatter, F., Liu, H.: Discovering, assessing, and mitigating data bias in social media. Online Social Networks and Media **1**, 113 (2017). https://doi.org/10.1016/j.osnem.2017.01.001
5. Morstatter, F., Pfeffer, J., Liu, H.: When is it biased? Proceedings of the 23rd International Conference on World Wide Web - WWW 14 Companion (2014). https://doi.org/10.1145/2567948.2576952
6. Morstatter, F., Pfeffer, J., Liu, H., Carley, K.M.: Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose (2013)
7. Summers, E., Summers, E.: On forgetting (Nov 2014), `https://medium.com/on-archivy/on-forgetting-e01a2b95272`
8. Wang, Y., Callan, J., Zheng, B.: Should we use the sample? analyzing datasets sampled from twitters stream api. ACM Transactions on the Web **9**(3), 123 (2015). https://doi.org/10.1145/2746366
9. Woodfield, K., Ahmed, W.: Chapter 4: Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges., vol. 2, p. 79107. Emerald (2017)
10. Zhang, H., Hill, S., Rothschild, D.: Addressing selection bias in event studies with general-purpose social media panels. Journal of Data and Information Quality **10**(1), 124 (2018). https://doi.org/10.1145/3185048
11. Zubiaga, A.: A longitudinal assessment of the persistence of twitter datasets (May 2018), `https://onlinelibrary.wiley.com/doi/10.1002/asi.24026`